Evidence Based Library and Information Practice

Evidence Summary

A Content Analysis of Google Scholar: Coverage Varies by Discipline and by Database

A review of:
Neuhaus, Chris, Ellen Neuhaus, Alan Asher, and Clint Wrede. "The Depth and Breadth of *Google Scholar*: An Empirical Study." portal: Libraries and the Academy 6.2 (Apr. 2006): 127-41.

Reviewed by:
Virginia Wilson
SHIRP Coordinator, Health Sciences Library, University of Saskatchewan
Saskatoon, Saskatchewan, Canada
E-mail: v_e_wilson@hotmail.com

Received: 30 November 2006 **Accepted:** 8 January 2007

Abstract

Objective – To ascertain the coverage by discipline, publication date, publication language, and upload frequency of the scholarly articles found in *Google Scholar*.

Design – Comparative content analyses.

Setting – Electronic information resources accessible via the internet (both freely accessible and for-fee databases).

Subjects – Forty-seven online databases and *Google Scholar*.

Methods – The study compared the content of 47 databases (21 Internet resources freely available to the general public; 26 restricted-access databases) covering a variety of subjects with the content of *Google Scholar*. Each database was assigned to one of the following discipline categories: business, education, humanities, science and medicine, social science, and multidisciplinary. From April through July 2005, researchers generated random samples of 50 article titles from each of the 47 databases and searched the titles on *Google Scholar* to determine inclusion.

Related studies were conducted for publication date and publication language analysis, and for the *Google Scholar* upload frequency study. For the publication date study, random samples from one database (*PsycINFO*) with a high degree of variability in *Google Scholar* coverage were searched for 1990, 2000, and 2004. For the publication language study, *Google Scholar* coverage of *PsycINFO* articles in English was compared to coverage of *PsycINFO* articles published

in non-English languages. For the upload frequency study, two databases chosen for their high degree of coverage (*BioMed Central* and *PubMed*) were monitored to determine how often the new content was uploaded to *Google Scholar*.

Main Results – This study revealed that content covered by *Google Scholar* varies greatly from database to database and from discipline to discipline. Of the 47 databases studied, coverage ranged from 6% to 100%. Mean and median values of coverage for all databases were both 60%. The mean discipline category scores varied from the humanities databases at 10% coverage, to the social sciences and education at 39% and 41% respectively, to science and medicine databases at 76% coverage. Mean coverage was 77% for the multidisciplinary databases. Mean coverage of open access journal databases was 95%, freely accessible databases had 84% mean coverage, and single publisher databases had 83% mean coverage.

The publication language study found a bias towards English language publications. As well, a publication date bias was found – coverage of earlier dates was not as thorough as coverage of more recent publications. In the upload frequency study, for *BioMed Central* and *PubMed* there appears to be an approximately 15-week delay in the uploading of new material to *Google Scholar*.

Conclusions – The results of this study serve to alert researchers and information professionals that *Google Scholar* (in beta test mode at the time of the study) has poor coverage in certain areas. To those with access to commercial databases, this serves as a cautionary tale. To those with a dearth of commercial databases, *Google Scholar* is a welcome site and can provide at least some information. The researchers state that the search engine itself could make future content studies unnecessary if it decides to make its content collection methodology transparent to users. Upload frequency, *Google Scholar's* linking services, the advanced search option, and the "cited by" feature could all be subjects of future studies. For its first year in operation, *Google Scholar* offers a broad range of discipline coverage with substantial depth in some areas. At the time of the study, *Google Scholar* was working with libraries and vendors to connect search results to library-licensed full text.

Commentary

Google Scholar has certainly evoked mixed reactions from library and information professionals since its appearance. This study is a revealing one, as *Google Scholar* does not release information as to what content is included and how it is chosen. There has been no other content analysis undertaken to this depth thus far on *Google Scholar*. However, there are some issues about the methodology of the study that call the results into question. For example, the investigators included 47 databases in their comparative analysis. They go to great lengths to describe how article title samples are randomly generated from these chosen databases for inclusion into the study. However, they do not describe – or even mention – how the 47 databases are chosen.

The researchers divide the 47 databases into disciplinary categories. This is a necessary step in order to determine coverage by discipline. However, while there are 15 databases in the science and medicine category, there is only one business database included. Other databases broken down by discipline include 3 in education, 5 in the humanities, 7 in the social sciences, and 16 multidisciplinary databases. The unevenness of the databases by discipline calls into question the validity of the results. In particular, generalizing about business

coverage based on one database does not give accurate results as to how business as a discipline is covered by *Google Scholar*. The inclusion of so many multidisciplinary databases is problematic, as they skew the "by discipline" results. The randomly generated article titles from the multidisciplinary databases presumably included articles from a variety of disciplines. The percentage of coverage in the humanities, for example, does not include the humanities articles that may have been generated from the multidisciplinary databases. And finally, assigning databases to disciplinary categories can be an interpretive exercise. For this study, should the *ATLA Religion Database* be categorized as Humanities? Or should it remain in Social Science, where the researchers included it? As well, *Library Literature* could have been classified as Social Science, whereas the researchers chose Education.

Additionally, it is interesting to note that in the coverage portion of the study, *PubMed* had 100% coverage in *Google Scholar*. In the upload testing portion of the study, it is clear that there is significant lag time in the uploading of recent material from *PubMed*. This discrepancy points to the inadequacy of the 50 article per database sample size. A larger sample size, particularly as some databases contain millions of citations, would have given a more accurate picture of actual coverage of any given database.

This study is useful to information practitioners in a provisionary way. It is good to know what kind of content is covered by *Google Scholar* and what the deficits might be when helping patrons to navigate *Google Scholar*. It would have been more useful had the article focused solely on the content inclusion analysis, and saved the publication date and publication language study and the upload frequency study results for another paper. Packing all the results into one relatively short paper did all of the studies a disservice. This article is usable for librarians across all sectors, and in particular information professionals who do not have access to a wide range of for-fee databases. However, *Google Scholar* must still be navigated carefully, as this study is certainly not the definitive answer as to what the database/search engine includes.